International Journal of
Innovative Management, Information & Production ISME International © 2015 ISSN 2185-5439
Volume **6**, Number **2**, June **2015** **pp.38-46**

EXAMINATION OF SEQUENTIAL PROBABILITY RATIO TESTS IN THE SETTING OF COMPUTERIZED CLASSIFICATION TESTS: A SIMULATION STUDY

MINGCHUAN HSIEH
National Academy for Educational Research, Taiwan
New Taipei City, 23703, Taiwan
hm7523@hotmail.com

ABSTRACT: *Sequential Probability Ratio Tests (SPRT) was first proposed by Wald (1947) as a quality control procedure in manufacturing. Many studies have been employed to investigate the usefulness of this method. This procedure was developed as a technique for making pass-fail or mastery-nonmastery decisions in computerized classification test. This study examines the efficiency and accuracy of SPRT procedure using simulation approach. The results show that SPRT procedure can be influenced by several factors: power of the test desired, indifference region of the masters and non-masters, pool quality and the constraints set for item exposure control. Conclusions and limitations are presented at the end of this study.*

Keywords: SPRT; Simulation Approach; Classification Test.

1. Introduction. Computer classification test (CCT) is frequently used for certification or licensure test. There are several potential benefits to implement the classification test in computer. For example, the mastery decision can be made more accurately. For paper-pencil test, students just need to answer a certain number of questions correctly, then they will be classified as masters or not. But for those students who just pass with the borderline score, actually we do not have much confidence to say that the examinees are actually master the test or not. With utilizing IRT techniques, CCT can control the classification error rate to meet the user's goal. On the other hand, the length of the test can be shortened. For the certification test, it usually covers a lot of content and requires the examinees to answer many items. By implementing the test in CCT, fewer items will be needed to classify people than the traditional paper-pencil test. Other advantages such as flexibility in scheduling the test, improvement of the test security, the easiness to administering new types of tests and collecting the data are also prominent. (Wainer, 2000)

Two main procedures have been proposed to make the mastery decisions: (1) Computerized Mastery Test (Lewis and Sheehan, 1990) and (2) Sequential probability ratio test. (Reckase, 1983; Wald, 1947). Spray et al.(1996) & Yi et al. (2001) compared these two procedures and their results indicated that the SPRT method resulted in a shorter average test length at the lower cutting point when the error rates of the two methods were closely matched. And they concluded that SPRT has the advantage over CMT procedure in

terms of classification accuracy and efficiency. In this study, we will only focus on the SPRT procedure.

SPRT was first proposed by Wald (1947) as a quality control procedure in manufacturing. Many studies have been employed to investigate the usefulness of this method (Bondarenko, 2010). This procedure was developed by Ferguson(1969), Reckase(1983) and Spray and Recase (1987, 1996) as a technique for making pass-fail or mastery-nonmastery decisions in computerized classification test. In the criterion reference testing situations, it is necessary to classify examinees between two hypotheses:

$$H_0 : \theta \le \theta_0 - \delta = \theta_1 \quad \text{vs} \quad H_1 : \theta \ge \theta_0 + \delta = \theta_2$$

where θ is the ability of an examinee, θ_0 is a cutting point, θ_1 and θ_2 refer to the lower bound and upper bounds. The width of $\theta_2 - \theta_1$ is called the indifference region, which it usually equals to 2δ. Two types of error are possible in making classification decisions: a false positive (α) or a false negative error (β). False positive is the examinee classified as a master but in fact his/her ability level is below the cutting point; on the other hand, if the examinee is classified as non-master but in fact his/her ability level is at or above the cutting point, a false negative error occurs. The relative importance of these two types of errors depends on the purpose of the test.

With SPRT, items are usually selected to maximize information at the cutting point, and this is also the most efficient selection way for dichotomous classification test (Beliler, 1998). Decisions are made not based on examinee's ability but on a likelihood ratio of alternative and null hypothesis, that is

$$LR(\underset{\sim}{x}) = \frac{\prod_{i=1}^{k} p_i(\theta_1)^{x_i}[1 - p_i(\theta_1)^{1-x_i}]}{\prod_{i=1}^{k} p_i(\theta_o)^{x_i}[1 - p_i(\theta_0)^{1-x_i}]}$$

Where K represents the number of items or the test length, x_i is the observed dichotomous item response which follows the Bernoulli distribution, and $P_i(\theta_0)$ and $P_i(\theta_1)$ are the probabilities of a correct response to item i, conditional on θ_0 and θ_1. The classification error rates, α and β can be determined before test administration because the upper and lower bounds of the likelihood ratio test are defined as functions of α and β. However, the actual observed error rates, α^* and β^*, may be lower from those predetermined values, where usually $\alpha^* \le \alpha/(1-\beta)$ and $\beta^* \le \beta/(1-\alpha)$. (Spray & Reckase, 1987) With the specified error rates, the decision rule used can be defined as follows (Wald, 1947):

Accept H_0 : determine the examinee is a nonmaster when

$$LR(\underset{\sim}{x}) \le (1-\beta)/\alpha$$

Accept H_1 : determine the examinee is a master when

$$LR(\underset{\sim}{x}) \ge (1-\beta)/\alpha$$

Another item is administered if

$$\beta/(1-\alpha) < LR(\underset{\sim}{x}) < (1-\beta)/\alpha$$

The item responses were given to the examinee sequentially until a classification decision is made.

In practice, the minimum and maximum test lengths are usually specified. If the decision cannot be made after the specified maximum of items, a force classification will be made: reject H_0 if LR(x) is greater than the midpoint of the interval $[\beta/(1-\alpha),(1-\beta)/\alpha]$; otherwise, accept the null hypothesis. (Lin et.al, 2000)

2. Purpose of the study. The purpose of this study was to examine the impact of the different factors on the test efficiency and classification errors using SPRT procedure for two category classification test. Four factors will be examined, including the power level, indifference region, item exposure control, and pool quality. The research questions for this study are:

1. When better quality pools are used in SPRT procedure, will it improve the test efficiency and reduce the classification error?
2. What are the effects on observed classification error and test efficiency when the predefined criteria (power and indifference region) are different?
3. When setting the item exposure control for each item, will it increase the classification error and extend the average number of test items required to make a classification decision?

3. Method. The item pool used in this study was from a certification test item pool which containing 5 categories and totally 791 items. Calibrations were made using the Bilog-MG computer program based on responses over 5000 examinees to each item. The 3PL IRT model was used.

Table 1 shows the characteristics for 791 items in the pool. For this pool, the average a-parameter is 0.523, which is a little lower than the usual real-world pool which Wang (1995) surveyed. The b-parameter has a wide range from 2.6 to -4.4. Note that the shortage of items at difficult level (b > 2), there are only 6 items; most items are cluster at b values from -2 to 0. Maybe it is because the passing theta for this test is usually set at -0.3

TABLE 1. Descriptive Statistics of Real Item Pool

Variable	N	Mean	Std	Minimum	Maximum
a-parameter	791	0.523	0.212	0.182	1.900
b-parameter	791	-0.853	1.133	-4.456	2.625
c-parameter	791	0.186	0.034	0.078	0.330
a-parameter (b > 2)	6	0.422	0.06	0.336	0.509
a-parameter (0 < b <2)	170	0.528	0.209	0.182	1.231
a-parameter (-2 < b < 0)	496	0.542	0.218	0.201	1.900
a-parameter (b < -2)	119	0.442	0.175	0.195	1.310

3.1. Simulation procedure. The simulation analyses were done using the SPRT Fortran program. The procedures are implemented as follows (1) Each simulee was randomly generated from the standard normal distribution N(0,1) at each cutting theta. (2) Based on the SPRT procedures, items were administered sequentially to the simulee and the response for a simulee was generated by comparing $P_i(\theta)$ to a random deviate (e.g. d) which were drawn from an uniform [0,1] distribution. If $P_i(\theta)>$ d, then the items will be scored as

correct, otherwise, the items will be scored as incorrect. Each of this run consisted of 5000 simulees, and different cutting points: -2, -1, 0, 1, 2 were investigated for each simulation.

The item information table for each cutting point is calculated in advance and sorted by information. Item selection is based on this table at each step until the decision can be made. For each simulation run, the initial item was randomly selected from those items at cutting theta equal to zero. Test will be ended when the likelihood ratio is below the lower bound (the examinees are classified as non-masters) or above the upper bound (the examinees are classified as masters). Besides, the minimum and maximum test length was specified as 80 and 100. When the decision cannot be made after 100 items, a force classification will be made.

3.2. Four simulation conditions.

1. Level of Power

Power is the probability that correctly reject the null hypothesis. That is, when we reject the null hypothesis and determine the examinee as a master, the true ability of this examinee is above the cutting point. Three levels of power: 95%, 70% and 40% were examined in this simulation condition.

2. Pool Quality

In addition to the real pool, two ideal pools were generated to compare the pool quality. For the first ideal item pool, 791 sets of parameters were generated from a pseudo-random generator with a-parameter distributed as N (2, 1), and b-parameter distributed as N (0, 2) and c is set at a constant 0.15. In this pool, the mean of the a-parameter is 2. This was called the high discrimination pool. The second ideal pool were also generated from the pseudo-random generator but with moderate a-parameter distributed as N (1, 1), and b-parameter distributed as N(0,2) and c = 0.15 for all items. The mean of a-parameter in this pool is 1 which is slightly lower than the mean of the high discrimination pool. So it was called the moderate discrimination pool.

In this simulation condition, three pools: high discrimination pool, moderate discrimination pool and real pool will be used as an independent variable to check the influence on test length and classification error rate.

3. Indifference Region

The indifference region is the distance between the masters and non-masters ($\theta_1 - \theta_0$). For this simulation condition, three sizes of indifference regions around cutting point were selected. The selected widths of the regions were 0.4, 0.6, and 0.8.

4. Item exposure control

For test security reasons, it is undesired to always select the best items in the pools. In this study, the best item refers to the items with the highest item information at the cutting point. Without item exposure control, the item overlap rate between any two test administrations will be very high and eventually lead to overexposure. To avoid this problem, some procedures were usually implemented during the item selection (Davey & Parshall, 1995; Kingsbury & Zara,1989; Sympson & Hetter, 1985)

The Sympson & Hetter (1985) procedure is a typical approach to controlling item exposure for CCT examinations. The algorithm was designed to put an upper bound on the exposure rate for each items and use it to control the item selection during the adaptive

testing. When selecting the next item, it will not only based on the item information at the cut point but also based on the item exposure control parameter. By setting a low exposure parameter for highly informative items, the overexposure problem can be prevented.

In this simulation run, the control parameter in S-H procedure will be examined. Four exposure control parameters r = 0.1 0.25, 0.4 will be compared with items without any item exposure control.

4. Simulation Result. The results from the SPRT simulation have been summarized in terms of test efficiency and decision accuracy. Efficiency was assessed by the average test length and accuracy by the classification error rate. These two outcomes are usually to be considered as the important indices to evaluate SPRT procedures.

4.1. Effect of power level. The results for each power level at each cutting point have been summarized in Figure 1 and Table 2. Figure 1 shows that when the power level gets higher, the test will take longer to make the classification decision, especially around the cutting point equal to zero. Table 2 shows observed false positive error and false negative error for each power level at each cutting point. It shows that power = 0.95 has the least false negative error rate. It can be explained that the power is the function of the negative error rate. (power = 1- β and the observed $\beta^* \leq \beta/(1-\alpha)$). Thus, with higher power, it yielded the lower observed false negative error rate. But for the observed false positive error rate, it is randomly distributed across three levels of power.

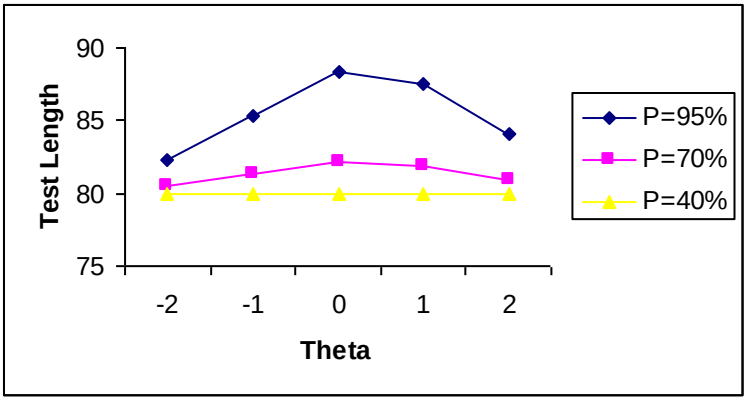

FIGURE1. Test length for different power level.

TABLE 2. Classification error rate for each power level.

Cutting Theta	Power =0.95		Power= 0.70		Power=0.40	
	False(+)	False(-)	False(+)	False(-)	False(+)	False(-)
-2	0.004	0.009	0.003	0.009	0.005	0.012
-1	0.019	0.026	0.024	0.029	0.023	0.033
0	0.040	0.040	0.045	0.041	0.044	0.048
1	0.040	0.025	0.034	0.027	0.037	0.026
2	0.017	0.004	0.016	0.004	0.023	0.004

Note: (1) Each mean is based on 5000 simulees. (2) The size of the indifference region around each passing score was set at 0.4. (3) The nominal error rate α = 0.05. (4) The target item exposure control was set at 0.2, using unconditional Sympson-Hetter.

4.2. Effect of indifference region. The result in Figure 2 shows that with the narrower indifference region, the longer test it will be, but this difference of test length are smaller at the extremes (cutting point = -2) Table 3 indicates that narrower indifference region generally yield lower false positive and false negative error rate.

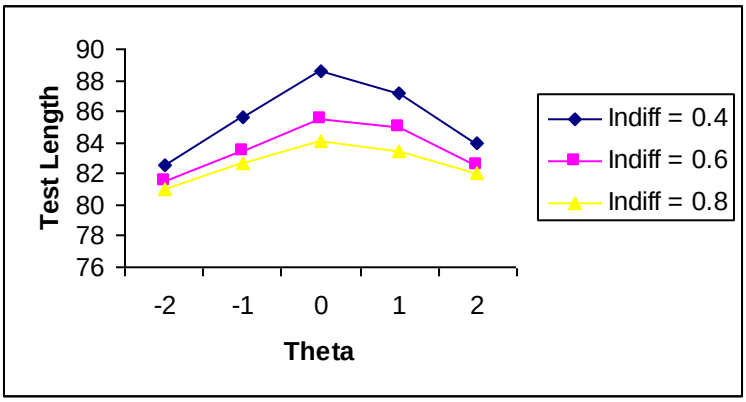

FIGURE 2. Test length for different indifference region.

TABLE 3. Classification error rate for each indifference region.

Cutting Theta	Indiff=0.4		Indiff= 0.6		Indiff=0.8	
	False(+)	False(-)	False(+)	False(-)	False(+)	False(-)
-2	0.005	0.011	0.005	0.012	0.004	0.011
-1	0.019	0.026	0.020	0.030	0.020	0.032
0	0.040	0.039	0.042	0.041	0.045	0.045
1	0.038	0.024	0.041	0.022	0.042	0.024
2	0.017	0.005	0.020	0.005	0.018	0.005

Note: (1) Each mean is based on 5000 simulees. (2) The nominal error rate α = β = 0.05 (3) The target item exposure control was set at 0.2, using unconditional Sympson-Hetter.

4.3. Effect of pool quality. Figure 3 and Table 4 show that the pool quality has some impact on the test efficiency and decision accuracy. When the items in the pool are highly discriminated, the test length can be shortened and the classification error can be reduced. For example, high discrimination pool needs the minimum items (80 items) among three pools to make the pass-fail decision, and it also yields the smallest false positive and false negative error rates for all cutting points.

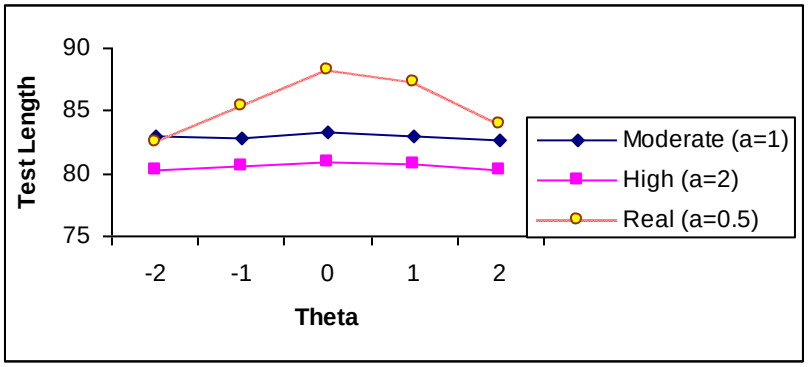

FIGURE 3. Test length for three pools.

TABLE 4. Classification error rate for three pools.

Cutting Theta	Real pool(a=0.5)		Moderate(a=1)		High(a=2)	
	False(+)	False(-)	False(+)	False(-)	False(+)	False(-)
-2	0.006	0.015	0.005	0.016	0.005	0.011
-1	0.020	0.026	0.012	0.017	0.009	0.011
0	0.040	0.039	0.020	0.019	0.017	0.013
1	0.038	0.026	0.023	0.017	0.011	0.008
2	0.015	0.006	0.020	0.005	0.005	0.005

Note: (1) Each mean is based on 5000 simulees. (2) The nominal error rate α = β = 0.05 (3) The target item exposure control was set at 0.2, using unconditional Sympson-Hetter. (4) The size of the indifference region around each passing score was set at 0.4.

4.4. Effect of item exposure control. The test length under each item exposure control parameter can be seen in Figure 4. The figure shows that for each cutting theta level, setting no item control will have shortest test length. With stricter item exposure control parameter, the longer test it will be. However, the differences of test length between each level are very close, especially for cutting theta at the extremes.

The classification error rate for each item exposure control parameter is presented in Table 5 as the percentage of simulees misclassified at each cutting point. It shows that items without any exposure control have smallest false positive and false negative error. With stricter item exposure control (r = 0.1), the false positive and false negative error would increase.

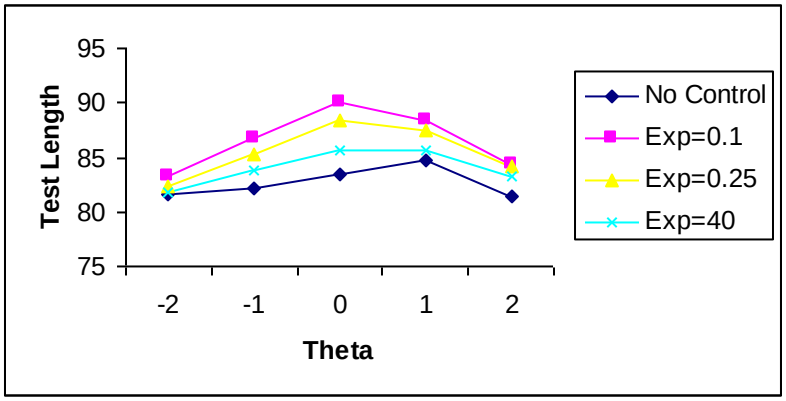

FIGURE 4. Test length for four level of item exposure.

TABLE 5. Classification error rate for four level of item exposure.

Cutting Theta	No control		Exp=0.1		Exp=0.25		Exp=0.40	
	False(+)	False(-)	False(+)	False(-)	False(+)	False(-)	False(+)	False(-)
-2	0.004	0.009	0.005	0.010	0.004	0.009	0.005	0.010
-1	0.016	0.019	0.025	0.035	0.023	0.026	0.022	0.024
0	0.027	0.026	0.049	0.041	0.040	0.042	0.038	0.032
1	0.029	0.020	0.049	0.032	0.036	0.025	0.035	0.023
2	0.016	0.005	0.017	0.005	0.017	0.007	0.016	0.005

Note: (1) Each mean is based on 5000 simulees. (2) The nominal error rate α = β = 0.05 (3) The size of the indifference region around each passing score was set at ± 0.2

5. Discussion. Wald (1947) has demonstrated that the simple sequential probability ratio test is the most efficient method of making classification decisions. Although in educational setting, it is much more complicated than Wald's analysis, some studies have proved that this procedure still have some advantages over other procedures (e.g. CMT procedure) in computer classification test. (Spray&Reckase, 1996; Yi et.al 2001).

However, the efficiency and accuracy of SPRT procedure can be greatly influenced by some factors: power of the test desired, indifference region of the masters and non-masters, pool quality and the constraints set for item exposure control. Based on the investigation of this study, four conclusions have been researched:

1. A higher power level would reduce the false negative error but it may increase the test length.
2. The larger the indifference region, the higher accuracy shorter the test will be, but it will trade off less decision accuracy.
3. The quality of the item pool has great impact on the SPRT procedure. Using high discrimination item pool can enhance the test efficiency and decision accuracy.
4. Setting the item exposure control for item selection will make the test longer and increase the classification error.

6. Limitation. Several limitations need to be addressed. First, the real item pool used in this study does not have good quality items and also have the shortage in higher theta level items; this problem may limit the generalization of this study.

Second, procedures for content balancing are not considered in this study. This may have unexpected results when combined this procedure with the item pool. Thus, it may be desirable to run more simulations and investigate the impact of the content balancing on test efficiency and decision accuracy.

Third, this study just focused on the two category decision test. It would be desirable to replicate this study with multiple cut point conditions.

REFERENCE

[1] Bleiler,T.L. (1998), *The precision of ability estimation methods in computerized adaptive testing*, Unpublished Doctoral Dissertation, The University of Iowa.

[2] Bondarenko, J. (2010), Sequential procedure for testing hypothesis about mean of latent gaussian process, *Applied Mathematical Science*, Vol.4, No. 62, pp.3083-3093.

[3] Davey,T., & Parshall,C.G. (1995), *New Algorithms for Item Selection and Exposure Control with Computerized Adaptive Testing*, Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

[4] Ferguson,R.L. (1969), *The Development, Implementation and Evaluation of A Computer-Assessted Branched Test for A Program of Individually Prescribed Instruction*, Unpublished Doctoral Dissertation, School of Education, University of Pittsburgh.

[5] Kingsbury, G.C., & Zara, A.R. (1989), Procedures for selecting items for computerized adaptive tests, *Applied Measurement in Education*, Vol. 2, No.4, pp. 359-375.

[6] Lewis, C., & Sheehan, K. (1990), Using Bayesian decision theory to design a computerized mastery test, *Applied Psychological Measurement*, Vol. 14, No. 4, pp. 367-368.

[7] Lin, C.J. & Spray, J.A. (2000), *Effects of Item-Selection Criteria on Classification Testing with the Sequential Probability Ratio Test*, Iowa city, IA: American College Testing Program.

[8] Reckase,M.D. (1983), A procedure for decision making using tailored testing, In D.J. Weiss (ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237-255). New York: Academic Press.

[9] Spray,J.A.& Reckase,M.D. (1987), *The Effect of Item Parameter Estimation Error on Decisions Made During The Sequential Probability Ratio Test*(Technical ONR 87-1), Iowa city, IA: American College Testing Program.

[10] Spray,J.A. & Reckase,M.D. (1994), *The selection of test items for decision making with a computer adaptive test*, Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

[11] Spray,J.A. & Reckase,M.D. (1996), Comparison of SPRT and sequential Bayes procedure for classifying examinees into two categories using a computerized test, *Journal of Behavioral and Educational Statistics*, Vol. 221, No. 4, pp. 405-414.

[12] Sympson, J. B., & Hetter, R.D. (1985), Controlling item exposure rates in computerized adaptive testing. *Proceedings of 27th Annual Meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

[13] Wainer H.(2000), *Computerized adaptive testing: a primer* (pp10-11) 2nd edition. NJ: Lawence Erlbaum Association.

[14] Wald,A. (1947), *Sequential analysis*. New York, NY: John Wiley and Sons.

[15] Wang,T. (1995), *The precision of ability estimation methods in computerized adaptive testing*. Unpublished doctoral dissertation, The University of Iowa.

[16] Urry,V.W. (1997), Tailored testing: A successful application of latent trait theory. *Journal Education Measurement*, Vol. 14, No. 2, pp. 181-196.

[17] Way,W.D. (1998), Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, pp. 17-27.

[18] Weiss,D.J., & Suhadolnik,D. (1982), Robustness of adaptive testing to multidimensionality. In D.J. Weiss (ed.), *Proceeding of the 1982 Item Response Theory and Computerized Adaptive Testing*, (pp.248-280). MN: The University of Minnesota.

[19] Yin Q, Handson B.A. Widiatmo H. Harris D.J. (2001), *Comparison of the SPRT and CMT procedures in computerized classification tests*. Paper presented at the Annual Meeting of American Educational Research Association, Seattle, WA.